AUTOMATION OF HEARING IMPAIRMENT TEST BASED ON AUTOMATIC SPEECH RECOGNITION

By

Captain Muhammad Danish Jillani

Captain Umar Farooq Malik

Captain Khushdil Abrar

Captain Adnan Haider

Supervised by:

Assistant Professor Dr Shibli Nisar

Submitted to the Faculty of Electrical Engineering, Military College of Signals
National University of Sciences and Technology, Rawalpindi in partial fulfillment
for the requirements of a B.E Degree in Telecommunication Engineering

June 2022

In the name of ALLAH, the Most Generous, the Most Courteous

CERTIFICATE OF CORRECTIONS AND APPROVAL

This is to officially state that the thesis work contained in this report

"Automation of Hearing Impairment Test Based on Automatic Speech Recognition"

is carried out by

Captain Muhammad Danish Jillani

Captain Umar Farooq Malik

Captain Khushdil Abrar

Captain Adnan Haider

under my supervision and that in my judgement, it is fully ample, in scope and excellence, for the degree of Bachelor of Electrical (Telecom.) Engineering in Military College of Signals, National University of Sciences and Technology (NUST), Islamabad. The material that has been used from other sources it has been properly acknowledged / referred.

Approved by

Asst. Prof. Dr. Shibli Nisar
(Supervisor)
Department of EE, MCS

Dated: 23 May, 2022

DECLARATION OF ORIGINALITY

We hereby declare that no portion of work presented in this thesis has been submitted in support of another award or qualification in either this institute or anywhere else.

ACKNOWLEDGEMENTS

We are thankful to Allah (the ALMIGHTY), to have guided us throughout this project, in every thick and thin. Without His invaluable assistance and guidance, we would have been unable to accomplish anything. We owe a debt of gratitude to our Parents & families for their unwavering moral support, which has helped us become who we are. We are extremely grateful to our project supervisor Asst Prof Dr. Shibli Nisar for his valuable time, patience, and efforts he has spent on us. We would also like to thank our instructors, professors, and all panelists in Military college of Signals (NUST) who taught us and helped us to complete our course work. We would also like to mention our friends and all the people who have helped us with our research.

PLAGIARISM CERTIFICATE (TURNITIN REPORT)

This thesis has **10%** similarity index. Turnitin report endorsed by Supervisor is attached.

Captain Muhammad Danish Jillani
NUST Serial no. 280744

Captain Adnan Haider
NUST Serial no. 280718

Captain Umar Farooq Malik
NUST Serial no. 280758

Captain Khushdil Abrar
NUST Serial no. 280750

Signature of Supervisor

ABSTRACT

Hearing impairment is found in 5% of the total world's population, out of which 10% are children and remaining 90% adults. If we focus on the age group of 65 years and above, one third of the group is facing hearing impairment issues. Different surveys carried out in USA revealed that Speech Recognition Threshold (SRT) Test is being conducted by 99.5% and 83% of audiologists for hearing assessment of the patients. However, not only the non-availability of expert audiologist but also the low literacy rate is a hurdle to conduct a successful Speech Recognition Threshold Test in Pakistan. A per the surveys Khyber Pakhtoon Kha (KPK) region of Pakistan has a literacy rate of 50%. Such less literacy rate along with Pushto language barrier between patient and audiologists makes hearing impairment diagnosis a troublesome process. This paper proposes an AHIT system based on SRT in Pushto language, which will be capable of detecting hearing impairment using Pushto as a test language, helping in automating the process of hearing impairment testing. The technique involves the extraction of Mel-Frequency Cepstral Coefficients (MFCCs) from a large data set of 15 different Pushto language spondee words. A Convolutional Neural Network (CNN) based machine learning model is then trained and tested using these MFCCs, whereas an interpreter code is used to test the system.

Table of Contents

List of Figures

List of Tables

Chapter 1: INTRODUCTION

1.1 Overview

Automated Hearing Impairment Testing (AHIT) is a Machine Learning based Automated Speech Recognition system, which enables to automate the process of Hearing Impairment Testing based on Speech Recognition Threshold (SRT) in Pushto Language. It provides an aid to the audiologist and patients of hearing impairment to perform the SRT test automatically in Pushto language.

1.2 Problem Statement

Hearing impairment is now found in 5% of the total world's population, out of which 10% are children and remaining 90% adults [1]. If we focus on the age group of 65 years and above, one third of the group suffers from hearing impairment [2]. Different surveys carried out in USA revealed that Speech Recognition Threshold (SRT) Test is being conducted by 99.5% [3] and 83% [4] of audiologists for hearing assessment of the patients. However, not only the non-availability of expert audiologist but also the low literacy rate is a hurdle to conduct a successful Speech Recognition Threshold Test in Pakistan [5]. A per the surveys Khyber Pakhtoon Kha (KPK) region of Pakistan has a literacy rate of 50% [6]. Such less literacy rate along with Pushto language barrier between patient and audiologists makes hearing impairment diagnosis a troublesome process.

1.3 Proposed Solution

Automated Hearing Impairment Testing (AHIT) aims at developing a prototype of the hearing impairment testing based on speech recognition threshold test, using machine learning, which will be capable of detecting hearing impairment using Pushto as a test language, helping in automating the process of hearing impairment testing.

1.4 Objectives

1.4.1 General Objectives

"An efficient establishment of an Automated Speech Recognition Threshold Testing in Pushto language, providing a hearing impairment testing device based on SRT to serve the humanity".

1.4.2 Academic Objectives

1.4.2.1 Development of an Automated Hearing Impairment Testing System.

1.4.2.2 To implement Machine Learning techniques and simulate the results.

1.4.2.3 To increase productivity by working in a team.

1.4.2.4 To design a project that contributes to the welfare of society in health sector.

1.5 Scope

The scope of our FYP is limited to spondee Pushto language words only. We will be covering limited words of Pushto language, which will be enough to train and deploy AHIT. In addition, initially we will be training AHIT for Pushto language only. Later, after the completion of FYP prototype, the system can also be upgraded to other local/international languages.

1.6 Deliverables

1.6.1 An ASR based system to identify hearing impairment by measuring SRT of the patient.

1.6.2 A publishable research paper on the topic.

1.7 Relevant Sustainable Development Goals

1.7.1 Good Health and Well Being

1.7.2 As many as 360 – 538 Million people across the world suffer from hearing loss as per WHO authorities [1]. Automation of Hearing Impairment Test will: -

1.7.2.1 Allow SRT measurement in Pushto language.

1.7.2.2 Reduce the Long-time appointment, consultancy fee, unavailability of equipment in remote areas.

1.7.2.3 Assist the audiologists by providing him the second option of hearing test other than PTA.

1.8 Structure of Thesis

1.8.1 Chapter 2 contains the literature review and the background and analysis study this thesis is based upon.

1.8.2 Chapter 3 Describes ASR and covers the functionality of the proposed solution.

1.8.3 Chapter 4 discusses the architecture, designing and development of AHIT.

1.8.4 Chapter 5 contains the analysis of the system.

1.8.5 Chapter 6 highlights the future work that is aimed to be produced.

1.8.6 Chapter 7 Conclusion

1.8.7 Chapter 8 Bibliography

Chapter 2: LITERATURE REVIEW

PART-1 SPEECH AND ASR

2.1 What is Speech?

Speech is formed by different waves that change air pressure, which can be apprehended through fervor of vocal cords. When air moves through vocal cords radiating acoustic energy from vocal tract, acoustic signals are generated. These acoustic signals then cause listener's eardrum to move in and out depending upon the pressure fluctuations, thus transforming the acoustic energy into mechanical energy at eardrum. When mechanical energy reaches as neural energy to the listener's brain, it is processed as sound. Human speech frequency is ~85Hz – 8kHz and hearing frequency is ~50Hz – 20kHz.

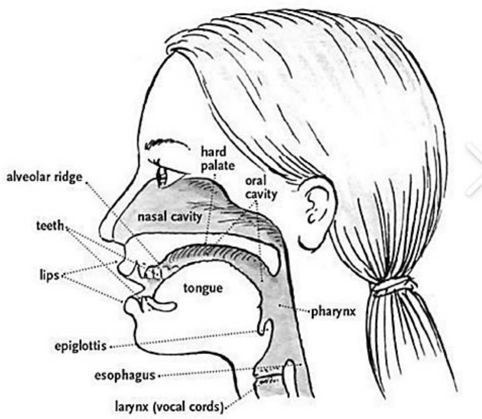

Figure 2.1 Speech signal basics

2.2 Automatic Speech Recognition (ASR) System

ASR is a branch of Artificial Intelligence that enables the translation and recognition of spoken speech into text. It can also be described as a technology that let human beings talk to a computer interface like a normal conversation between human beings. ASR Systems are serving in many ways to automate different speech related processes. These are mainly divided into two different streams i.e., one is dependent on speaker and the other is speaker independent system. Speaker dependent systems are those in which system will only respond to the speakers whose voice is enrolled in it as training data. This is done by making speaker read some text or isolated words. While in case of speaker independent speech recognition system, the system does not rely on vocal training and responds to every speaker.

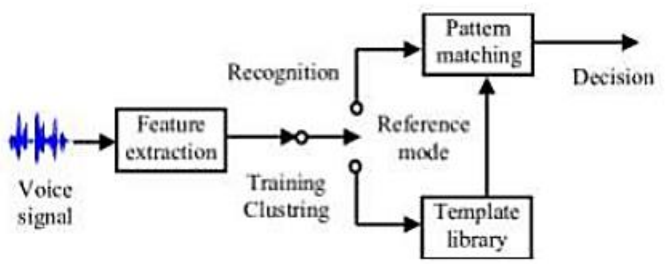

Figure 2.2 General ASR System

2.3 Definition of Natural Language Processing (NLP)

NLP is an advanced version of the ASR systems. It allows human to interact with computer through their voices and text. We humans interact with each through both these forms i.e., written text and voices. Therefore, to make our lives easier, we need to invent ways to make our computers understand us in both these forms. In addition, NLP is the solution to all of this.

Study of NLP has been around for 5 decades now, and it is still growing. We can clearly see the results in form of our smartphones responding to our voices as Siri or Bixby. Accuracy of around 96 to 99% has been achieved in different NLP programs. It is programmed on a very large vocabulary and as it is almost impractical to scan the whole vocabulary hence it is designed to mark the tagged words only and respond to them.

2.4 How does an ASR System Works?

As a first step noise is removed from the audio data, followed by features extraction using Mel-Frequency Cepstral Coefficient (MFCC). To extract features, audio needs to be passed through few steps as represented in the block diagram below.

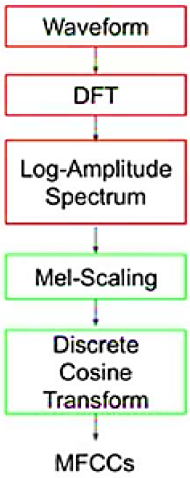

Figure 2.3 MFCC extraction

The speech signals are firstly converted from analogue to digital signals. It is because computer cannot process on analogue data. This is done by Analogue to Digital Converter (ADC) in three steps i.e., sampling, quantization and encoding.

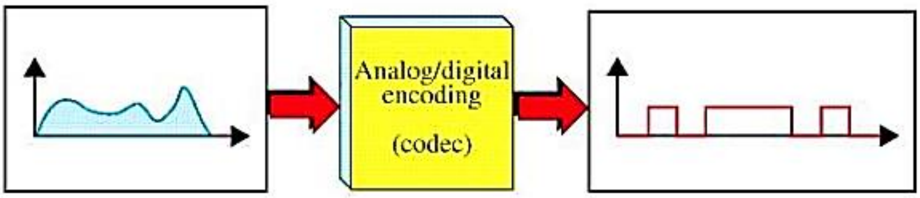

Figure 2.4 Analog to digital

Once the data has been converted into digital format, we then apply Fast Fourier Transform (FFT) to convert the graph in spectrogram. Vertical axis of the spectrogram displays frequency, and horizontal axis shows time and color shows the intensity that makes the sound. The lighter the color the more energy is used.

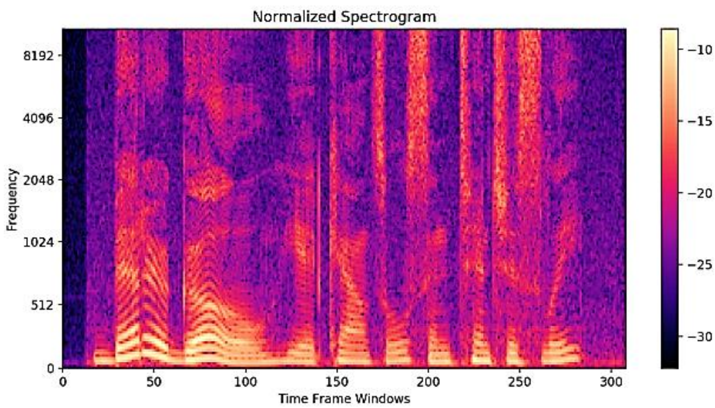

Figure 2.5 Spectrogram

After it is done the features of data can be extracted. Usually, first 12 to 13 coefficient are considered most important because they contain most of the important information like formants and spectral envelopes etc. In total there are 39 coefficients per frame

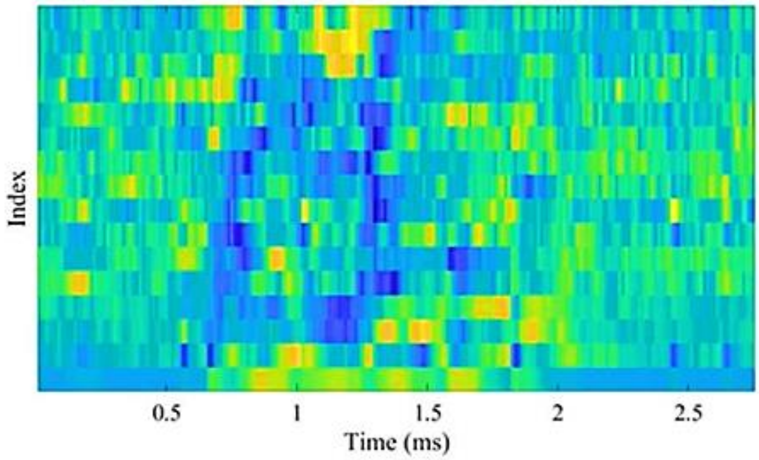

Figure 2.6 Frames

After these features are extracted, the required algorithm is applied depending on the system.

2.5 Problems faced in making as ASR system

With world moving towards this technology for their ease, following are the major problems which are faced in deploying such systems in real life everywhere. The biggest of all problems in noise.

Noise is everywhere around us and it disrupts the sound waves making it difficult for the computer to understand and differentiate between different words.

Accent is another huge barrier in enhancing the efficiency of ASR systems. Every language has multiple accents and if a word is pronounced in different ways, then the phonemes for each accent will vary making it difficult for an ASR system to process.

Moreover, echo and similar sounds of words is also a problem but apart from all that the speed by which this technology is enhancing, all big companies are working on making their own ASR systems to enable automation of all tasks.

PART-2 SPEECH RECOGNITION THRESHOLD (SRT)

2.6 Different Thresholds for Speech

2.6.1 Threshold of a non-speech signal

The threshold of a nonspeech signal such as a tone has a clear meaning: it is the level at which the presence of the sound is just audible. However, the threshold for speech can mean the lowest level at which speech is either just audible or just intelligible.

2.6.2 Speech Detection or Awareness Threshold (SDT/SAT)

The lowest level at which the presence of a speech signal can be heard 50% of the time is called the speech detection threshold (SDT) or the speech awareness threshold (SAT).

2.6.3 Speech Recognition / Reception Threshold (SRT)

The speech recognition threshold or speech reception threshold is the lowest level at which a voice signal is intelligible enough to be identified or recognized 50% of the time (SRT). Spondee (or spondaic) words, which are two-syllable words with equal stress on each syllable, such as "baseball" or "railroad," are often used to obtain the SRT. Spondaic words work well for this since even just a minor increase in the intensity causes spondee recognition to jump from 0% to 100% in a matter of a few seconds. When evaluated with spondaic words, the SRT is also known as the spondee threshold (ST) [16].

2.6.4 SDT vs SRT

Because the SDT relies solely on audibility, the SRT requires inducements to be heard and recognized, the speech detection and recognition thresholds must be different. The finding that the average SDT is roughly 7 to 9 dB lower (better) than the mean SRT [16] has repeatedly confirmed this hypothesis. As a result, the average SDT-SRT difference for a male talker is 8 dB and 8.2 dB for a female talker.

2.7 Clinical Functions/Purpose of SRT Test

The speech reception threshold serves a variety of medical application :-

2.7.1 To be used as an evaluation tool for pure tone thresholds.

2.7.2 To serve as a guideline for determining appropriate levels for administering suprathreshold speech recognition tests.

2.7.3 To determine the requirements and performance of hearing aids.

2.7.4 To determine necessity of auditory (re)habilitation and management method refinement.

2.7.5 To determine hearing sensitivity in small children and others who are difficult to test.

2.7.6 To describe the level of hearing loss and how it affects speech comprehension.

2.8 Methods for SRT Testing

The SRT is widely considered to be the lowest audible level at which a patient can repeat 50% of spondee syllables, however there are many other methods for determining that point. Although their individual properties can vary greatly, most SRT testing methods share several basic characteristics. The most common element is that the patient is shown multiple spondee words one by one at the same audible level.

2.8.1 The Descending Method

The descending methods begin presenting these blocks of test words above the estimated SRT so that the patient is initially able to repeat them, and then present subsequent blocks of spondee words at progressively lower hearing levels. This process is repeated until the patient misses a certain number of words; at which time the descending run is over.

2.8.2 The Ascending Method

The ascending methods start below the estimated threshold, where the patient cannot repeat the words, and then present subsequent blocks of test words at progressively higher hearing levels. This procedure is repeated until the patient correctly repeats a certain number of words, at which point the ascending run is terminated. Features such as the number of words in a block, the criteria for starting and ending an ascending or descending run, and how the "50 percent accurate" threshold is defined distinguish the approaches.

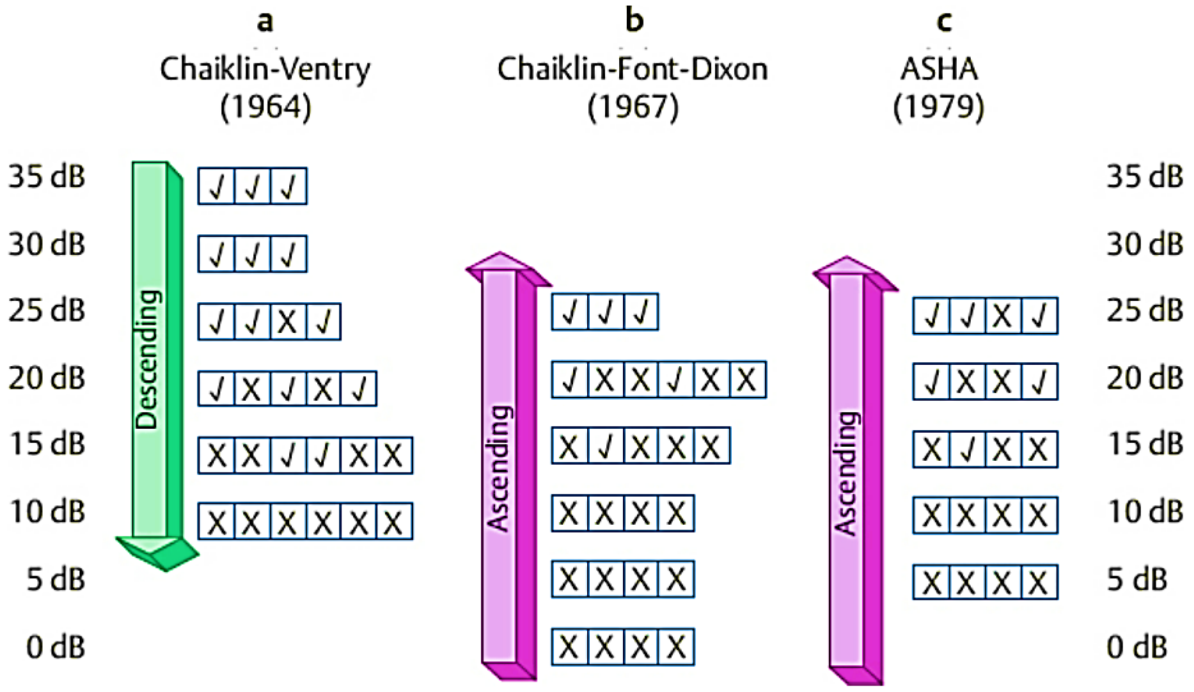

Figure 2.7. The SRT can be measured in a variety of ways, including descending (Chaiklin & Ventry 1964), ascending (Chaiklin, Font, & Dixon 1967), and ascending (ASHA 1979). [16]

2.9 Clinical assessment from SRT

The ability to interpret spoken language and the threshold for speech are easily distinguishable. Listen to the following complaints from different hearing-impaired patients: "I can hear but not understand speech"; "Words aren't clear"; "Speech sounds muffled [or twisted]"; "I mix up words." The most prevalent notice in this situation is that the patient's speech is lacking in fluidity. This issue manifests itself in the form of received messages that are incorrect and unclear. A patient exact diagnosis can be performed by the audiologist by knowing the sequence and pattern of SRT.

2.10 Previous Work

This section covers the previous work done on ASR systems and on Pashto dataset formation. In the recent decade, a huge amount of research was carried on local languages including Urdu, Punjabi, Hindi, and Arabic.

2.10.1 Speaker Independent Urdu Speech Recognition using HMM

Javed Ashraf et al [7]**,** presented an ASR system for limited vocabulary that was speaker independent. This system works for isolated words only. The system is based on Sphinx4 and HMM having a word error rate of 10%.

Sphinx-4 framework can function efficiently for small to midsized vocabulary, according to the study findings. The acoustic model for this system was created entirely by the project's authors.

Speaker	WER (%)			
	Test 1	Test 2	Test 3	Mean WER
Speaker 1	10	10	10	10.00
Speaker 2	10	10	20	13.33
Speaker 3	10	0	10	6.66
Speaker 4	20	10	10	13.33
Speaker 5	10	10	10	10.00
			Mean	10.66

Table 2.1 Research WER [7]

2.10.2 Development of the MIT ASR system for the 2016 Arabic multi-genre broadcast challenge

Tuka Al Hanai et al [8] described a 1,200-hour speech corpus, which was used to design an Arabic ASR system. A range of Deep Neural Network (DNN) topologies were analyzed, and was shown that a sequence discriminatively trained G-LSTM neural network had the best performance in terms of the time.. CNN model was found good at feature extraction representation and reducing variance in frequency domain. Thus, better results could be concluded from it if it was used within a hybrid-like DNN topology. The best overall Word Error Rate was 18.3% ($p < 0.001$). The significance of the results (8/13 results with $p < 0.05$ compared to next increase in WER) highlights that each incremental improvement in WER introduced by a different network topology is a significant increase, even if it is a difference of only 0.3% absolute.

Model	Topology	Features	Alignments	WER (%) p<(prev/base)	WER(%) 4gram p<(prev/base)
GMM-HMM	-	MFCC+LDA+MLLT +FMLLR	-	40.3(-/-)	-
DNN CE	5x1024	30 Fbank+Pitch	GMM	29.7(0.001/0.001)	28.1(0.001/0.001)
CNN	4x2000	80 Fbank+Pitch	GMM	29.5(0.472/0.001)	28.1(0.734/0.001)
TDNN	6x3000	80 Fbank+Pitch	GMM	27.1(0.001/0.001)	25.8(0.001/0.001)

DNN MPE	5x1024	30 Fbank+Pitch	CE	25.6(0.001/0.001)	24.7(0.001/0.001)
Chain TDNN	7x625	80 Fbank+Pitch	GMM	23.6(0.001/0.001)	23.4(0.001/0.001)
LSTM	3x1024	80 Fbank+Pitch	CE	23.6(0.936/0.001)	22.7(0.001/0.001)
H-LSTM 3L	3x1024	80 Fbank+Pitch	CE	23.3(0.027/0.001)	22.6(0.250/0.001)
H-LSTM 5L	5x1024	80 Fbank+Pitch	CE	23.1(0.055/0.001)	22.4(0.184/0.001)
G-LSTM 3L	3x1024	80 Fbank+Pitch	CE	22.4(0.001/0.001)	21.7(0.001/0.001)
G-LSTM 5L	5x1024	80 Fbank+Pitch	CE	22.2(0.110/0.001)	21.5(0.070/0.001)
G-LSTM 3L sMBR	3x1024	80 Fbank+Pitch	CE	20.4(0.001/0.001)	19.5(0.001/0.001)
G-LSTM 5L sMER	5x1024	80 Fbank+Pitch	CE	20.1(0.009/0.001)	19.2(0.034/0.001)
Top 2 Combined	G-LSTM sMBR (3L+5L)	80 Fbank+Pitch	CE	**-**	**18.3(0.001/0.001)**

Table 2.2 Results [8]

2.10.3 An ASR system for spontaneous Punjabi speech corpus

An ASR system for spontaneous Punjabi speech corpus was designed by Yogesh Kumar et al [9]. Capability of the system was to recognize live spontaneous speech in Punjabi language. The user interfaces were created with java programming language. To train the system 6012 Punjabi words while 1433 Punjabi sentences were used. The main intention of this study was to reduce the % error in the speech model.

ਸ	ਡ	ਤ	ਰ	ਨ	ਹ	ਲ਼	ਕ
ਪ	ਜ	ਦ	ਠ	ਅੰ	ਆ	ਪਿੰ	ਕ੍ਰਿ
ਮੈ	ਹਾ	ੳ	ਅ	ਬ	ਮ	ਨਾ	ਯ
ਗ	ਸ਼	ਸਿੱ	ਖਿ	ਕੰ	ਪਿ	ਊ	ਇੱ
ਨਿੰ	ਐ	ਈ	ਵਿ	ਚ	ਟ	ਕਿੰ	ਪੁ
ਸੀਂ	ਖ	ਉ	ਘ	ਤੁ.	ਨੂੰ	ਥ	ਤ੍ਰੇ
ਦਿ	ਛਿ	ਰਾਂ	ਵ	ਪੰ	ੁ	ੇ	ਾ
ੰ	ੈ	ੀ	ੋ	ਂ	ੱ	ੂ	!

Table 2.3 Phonetic files of Punjabi language

This accuracy of this system was examined during a Punjabi interview. Multiple Punjabi words and phrases were spoken. In terms of recognition rate, Punjabi phrases earned 90.8 percent, whereas Punjabi words got 93.79 percent. Detail is shown in the table 2.4.

Punjabi Interview Speech Corpus	**Correct**	**Error**	**Accuracy (Correct %age)**
Total No. of Punjabi Sentences = 461	455	6	98.6%
Total No. of Punjabi words = 1227	1213	14	98.8%

Table 2.4 Results [9]

2.10.4 Pashto Spoken Digits Database for the Automatic Speech Recognition Research

Recently a little work was also done on the Pashto language, but it was restricted to isolated digits only. Arbab Waseem et al [10], presented the advancement of the first Pashto isolated Spoken Digits database for ASR research consisting of digits from zero (sefer) to hundred (sul) voiced by 60 speakers, 30 male whereas 30 female. The attendees were between the ages of 18 and 60. The recordings were made with a Sony PCM-M 10 Linear Recorder in a noise-free situation. Mel Frequency Cepstral Coefficients (MFCC) were utilized as a feature, and a Linear Discriminant Analysis (LDA)-based classifier was used to classify the data. In terms of recognition rate, the performance was found to be 67 percent.

2.10.5 The Development of Isolated Words Pashto Automatic Speech Recognition System

Some other relevant work was done by Irfan Ahmed et al [11] in the Pashto language with a medium-sized corpus. The collection contains 50 utterances of 161 isolated most frequently spoken Pashtoo words, names of days of the week, and numbers ranging from 0 to 25. The dataset contains a recording from both genders i.e. 28 male and 22 female speakers having ages ranging from 16 to 40. The sampling frequency was 44100 Hz. The word error rate was calculated to be 33.33%.

Word Number	Percentage Error
1	0
2	33.33
3	0
4	60
5	33.33
6	33.33
7	50
8	33.33
9	33.33
10	33.33

Table 2.5 Results [11]

2.10.6 Pashto isolated digits recognition DCNN

Bakht Zada et al [12] revealed Pashto isolated digits recognition using deep convolutional neural networks, with superior results than earlier similar efforts. Each Pashto digit from 0 to 9 had 50 utterances in the vocabulary. For each isolated digit, twenty MFCC patterns were retrieved and fed into CNN as input. Testing revealed a total average accuracy of 84.17 percent, which is considerably higher than the previous solutions. Some more work done for Pashto is shown in [13][14].

Data	No. of Speakers	Accuracy
Training	25 (both male and female)	90.14%
Testing	25 (both male and female)	84.17%

Table 2.6 Dataset [12]

Research P#	Language	Algo	Accuracy	WER
1	Urdu	Sphinx4	90%	10%
2	Arabic	DNN	81.7%	18.3%
3	Punjabi	CNN	90.8%	9.2%
4	Pashto	LDA	67%	33%
5	Pashto	DNN	66.67%	33.33%
6	Pashto	CNN	84.17%	15.83%

Table 2.7 Research Analysis [12]

2.10.7 Speech Reception Threshold Measurement Using Automatic Speech Recognition

Emre Ylmaz [15] introduced this research concept. Different task-dependent language models are discussed in the works. Because the phrases were known in advance, a Finite State Grammar (FSG) was developed, enabling FSGs to be used for this recognition challenge. Listeners were evaluated solely based on distinct keywords in the statement, which might be repeated in any order. Non-keywords, on the other hand, can be skipped, introduced, or replaced.

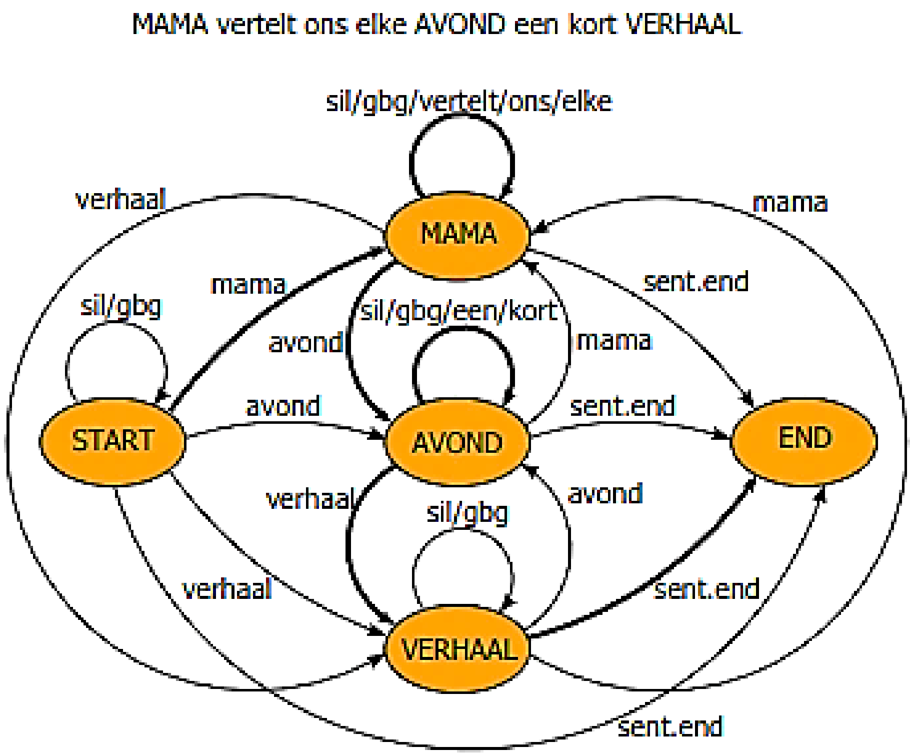

Figure 2.8. Finite state grammar (FSG)

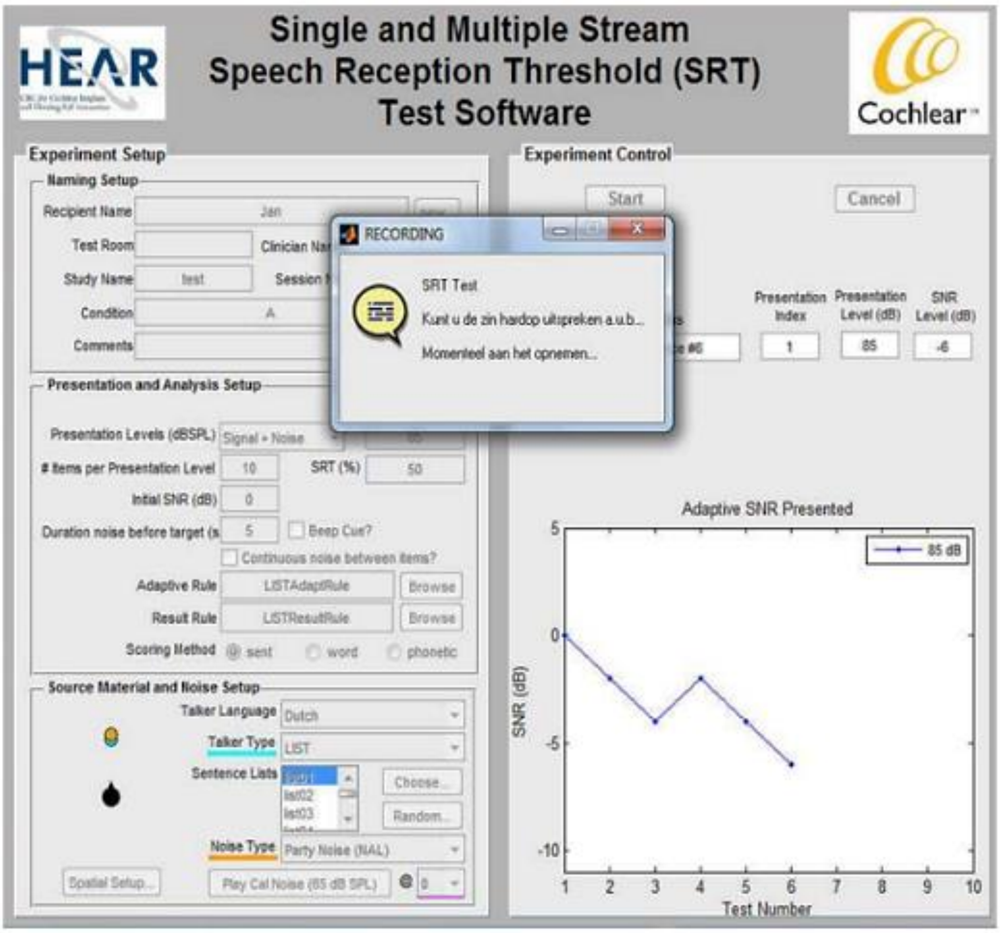

Figure 2.9. SRT Test Software

Chapter 3: FUNCTIONALITY OF PROPOSED SOLUTION

3 The system was designed in a way that the whole process is as smooth and precise as possible, and for this reason, we have mentioned the whole methodology in Figure 3.1. Our speech signal after the initial processing is given to the system as an input to be further inspected. Firstly, the feature extraction process is carried from the speech signal which will be explained below. For the recognition, we used the CNN model for our system as a priority. This was used to train our model. The training was carried on almost 80 % of our data and then accuracy was determined for respective Machine Learning Algorithms. The output of the system was then shown in the form of correctly / incorrectly recognized word along with SRT.

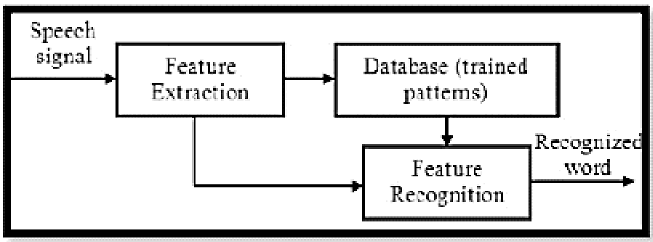

Figure 3.1 Block Diagram

3.1 MFCC Feature Extraction

3.1.1 What are MFCCs

Mel Frequency Cepstral Coefficients, Mel – frequency represents the implementation of Mel Spectrograms. The MFCC provides numerous coefficients defining various characteristics. In MFFC, the noun Cepstrum is derived from spectrum. The cepstrum domain derives quefrequency, liftering and rhamonic from frequency, filtering, and harmonic respectively.

Figure 3.2 Cepstrum Domain

Historically researchers at MIT came out with the concept of cepstrum while studying echoes in seismic signals (1960s). The concept was afterwards used for the audio feature of choice for speech recognition / identification (1970s). In 2000s Cepstrum started to be adopted in MFCCs specifically for Music processing.

Mathematically Cepstrum is written as $C(x(t)) = F^{-1}[log(F[x(t)])]$. Where F[x(t)] is the Fourier transform representing the spectrum of the time domain signal. Log function provides the logarithmic spectrum allowing to separate the components of the signals using logarithmic properties. In the end the inverse Fourier transform of the log spectrum provides the Cepstrum. The below graphic representation of the Cepstrum is shown in Figure 3.3.

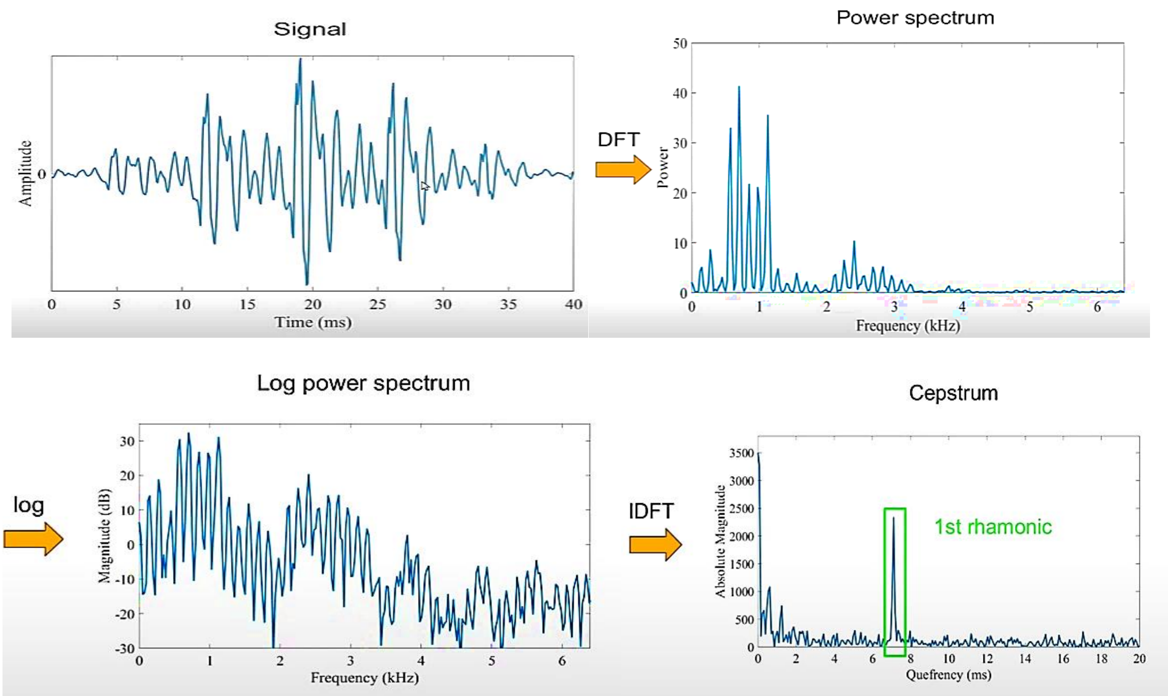

Figure 3.3 Graphic Representation of Cepstrum

Now for further information on MFCCs, we need to understand how speech works. The key element is the vocal tract that produces a sound. Depending on how we shape our vocal tract will produce various sounds. In other words, vocal tract acts as a filter. Hence, we can say that a speech signal is a convolution of signal of glottal pulse and vocal tract.

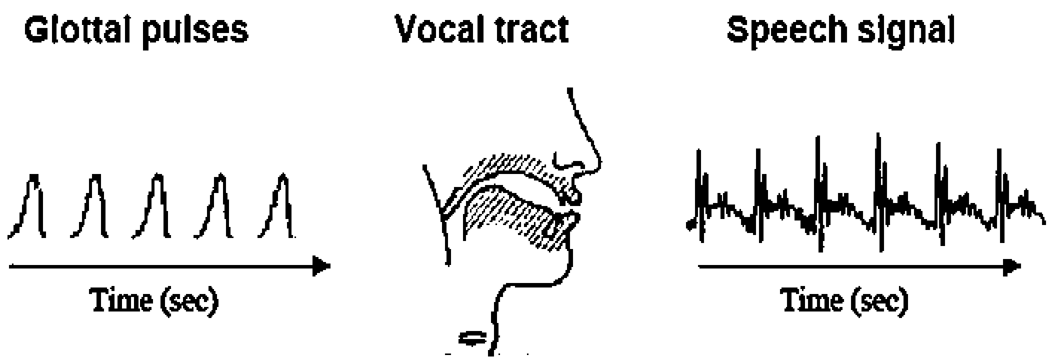

Figure 3.4 Sound (speech signal) production by humans

A spectral envelope is acquired from the speech log spectrum, which provides us Formants carrying different identities of the sound. The spectral details match with the response of glottal pulse signal being generated. As a result, speech can be seen as a complication of the frequency response of the verbal tract with the glottal pulse.

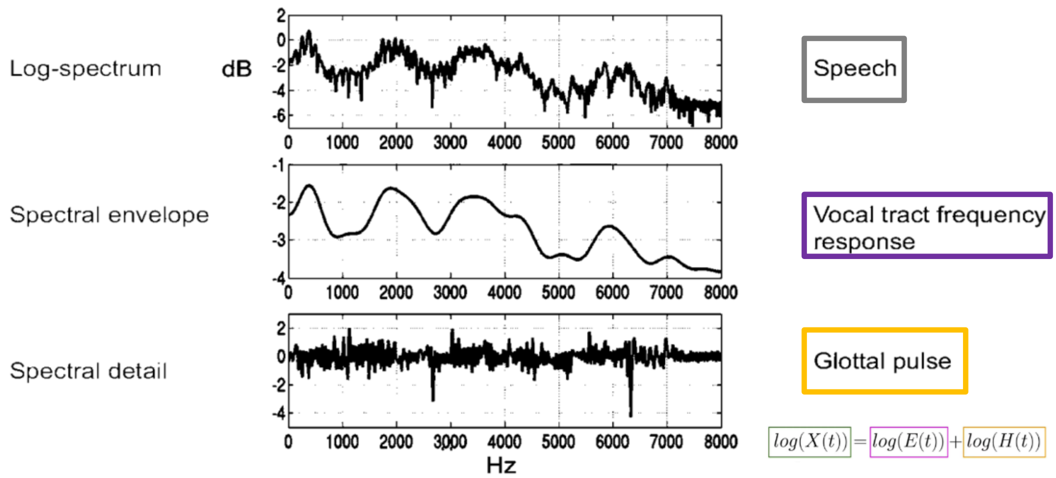

Figure 3.5 Speech is a convolution of vocal tract freq response and glottal pulse

3.1.2 MFCCs Calculations

Mel Frequency Cepstral Coefficients calculation begins with a DFT of waveform followed by logarithm to get log-amplitude spectrum. Mel Scaling is performed followed by Discrete Cosine Transform as it provides real valued coefficients called MFCCs.

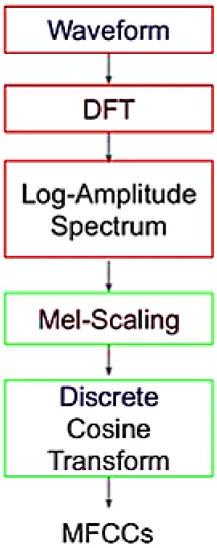

Figure 3.6 MFCCs Calculation

Traditionally first 12-13 coefficients are considered as they are the ones having the maximum information about the spectral envelope or vocal tract frequency response i.e. the formants.

3.2 CNN

Convolutional neural networks are machine learning-based models which works on a mathematical operation of convolution. Here we use convolution rather than the general multiplication of matrices in one of the layers. This neural network has different layers labeled as the input, output, and hidden layer. The middle layer is the one that is the hidden one because the activation function and the convolution mask the inputs and outputs of this layer. These hidden layers undertake the process of convolution. Here dot product is taken between the inputs of the layer i.e., in our case the MFCC features and the convolutional kernel. This kernel is then slid across the input matrix and generates a feature map which is deemed as the response for the following layer. Further, the processes of pooling and normalization layer etc. are performed. Along with the convolutional layer, there is also a pooling layer that reduces the data by combining the data clusters of the output of one layer to a distinct cluster for the upcoming layer. The two most frequent methods of pooling are max pooling and average pooling, in which the maximum value of each cluster and the average value are used, respectively. The next step is dropout where a neural network is prevented from overfitting and for this purpose, different probability values are tested for better accuracy. We will be discussing the results for our dataset while the convolutional neural network was applied to it in the next sections.

3.3 SRT

SRT (Speech Reception Threshold) is the lowest threshold at which a speech signal is recognized or identified 50% of the time. SRT is typically gained by asking patients to repeat spondee (or spondaic) words with equal emphasis on both syllables. Spondaic words work well for this since even a minor increase in intensity causes spondee recognition to jump from 0% to 100% in a matter of seconds.

Chapter 4: DESIGN AND DEVELOPMENT OF AHIT

4 Purpose

The goal of this chapter is to provide a comprehensive description of the Automated Hearing Impairment Testing system in Pashtoo, helping in automating the process of hearing impairment by audiologists, to bridge the communication gap between audiologists and the patients. The aim of this document is to present a detailed description of the project AHIT.

4.1 Overall Description

4.1.1 Product Functions

The main features of AHIT in this domain are highlighted below:

- The system plays randomly 6 x syllable words.
- User will respond to the voice heard.
- The sound will be compared by the trained CNN model.
- SRT algorithm is utilized to calculate the SRT.
- Results are displayed on the screen.

4.1.2 User Classes and Characteristics

The following section describes the types of users of AHIT.

4.1.2.1 **Audiologists**

Audiologists will have the access to the system who will be able to even extend the dataset according to their needs and resources once the prototype is in their hands.

4.1.2.2 **Patients**

Different patients will be able to check their hearing ability using the the device in their local dialect.

4.2 Dataset Collection

15 x Syllable words were selected in Pushto language. For each word various voice samples of different age groups and different dialects were taken. Almost 260 samples for each word and 260 x 15 = 3900 samples overall were collected. These words are as follows: -

Ser	Pashto	Meaning in English
1.	Khandaa	Smile
2.	Kitaab	Book
3.	Margai	Bird
4.	Karga	Crow
5.	Qisa	Story
6.	Saba	Tomorrow
7.	Topak	Gun
8.	Lota	Pitcher
9.	Qainchi	Scissors
10.	Qalam	Pen
11.	Poza	Nose
12.	Nika	Grandfather
13.	Carga	Hen
14.	Zalmai	Adult boy
15.	Oorbal	Head Hair

Table 4.1 Syllable Words for Pashtoo Language

4.3 Feature Extraction

As discussed in chapter 3.1 MFCCs are extracted from the audios as they carry the formants of the speech signal. To extract MFCCs, librosa library is used in python – jupyter Notebook. 26 x MFCCs are extracted and allotted against each audio sample. Following is the. json view of the allotted MFCCs.

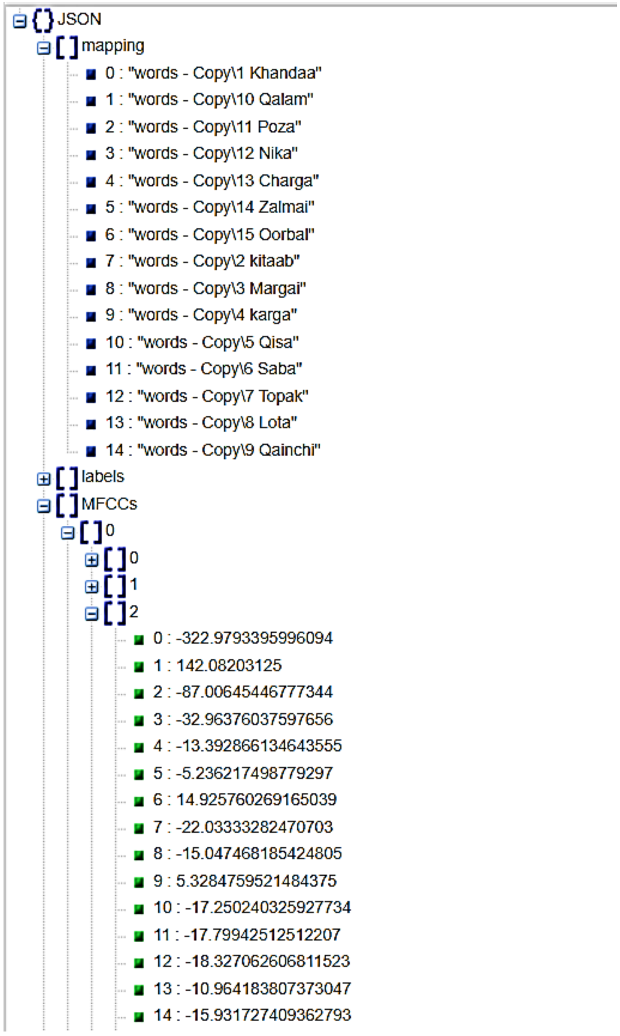

Figure 4.1 Extracted MFCCs

4.4 Model training

As per the discussion in Chapter 3.2, CNN model is employed to train the model on the basis of the features extracted from the data. Data is used for training in 80% of the cases and testing in 20% of the cases.

4.5 SRT Calculation

Methods for calculating SRT are already discussed in chapter 2.9. Here is the flow chart and algorithm of the descending SRT calculation method being employed.

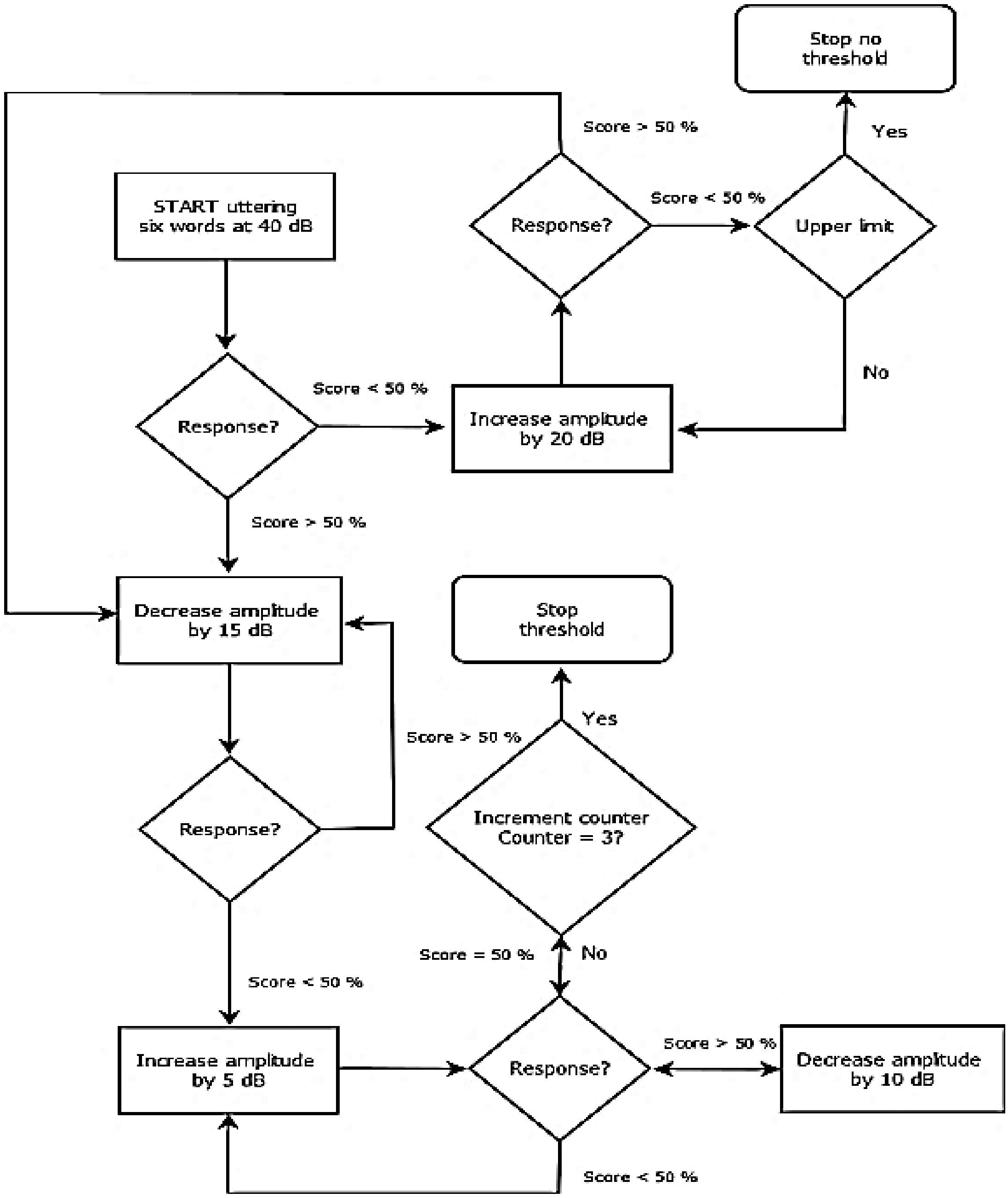

Figure 4.2 Flow chart of SRT calculation[17]

```
Require: Recorded spondee words, counter=0,
    status[3]=0, upper limit = 120 dB, A=45 dB
 1: procedure
 2:     Play(spondee, A)
 3:     if score > 50 then
 4:         while score > 50 do
 5:             Decrease amplitude by 15 dB
 6:             A=A-15
 7:         end while
 8:         while score =< 50 do
 9:             Increase amplitude by 5 dB
10:             A=A+5
11:         end while
12:         if score = 50 then
13:             counter++
14:             status[j++]=A
15:             if counter ==3 then
16:                 SRT = median(status[3])
17:                 break END TEST
18:             else
19:                 play(spondee,A)
20:                 go to step 12
21:             end if
22:         else
23:         end if
24:         if score ¿ 50 then
25:             Decrease tone by 10 dB,
26:             A=A-10
27:             play(spondee,A)
28:             go to step 12
29:         else
30:             Increase tone by 5 dB
31:             A=A+5
32:             play(spondee,A)
33:             go to step 12
34:         end if
35:     else
36:         while s docore =< 50 && threshold < upper
            limit do
37:             Increase amplitude by 20 dB
38:         end while
39:         if score > 50 then
40:             go to step 4
41:         else threshold >= upper limit
42:             No threshold found
43:         end if
44:     end if
45: end procedure
46: return Speech Recognition Threshold
```

Figure 4.3 SRT Calculation – Algorithm[17]

4.6 Operating Environment

4.6.1 Hardware

AHIT will have following hardware specifications:

- **Raspberry Pie 4:** For processing sounds as a standalone device.

- **High quality professional condenser microphone:** For recording the dataset/calls manually.
- **LCD**: For display interface of device.

4.6.2 Software

AHIT will have following Software specifications:

- Linux/Raspbian.
- Python IDE.

4.7 Design and implementation Constraints

4.7.1 The system requires a completely noise free environment [16].

4.7.2 It will be able to process one user at a time.

4.7.3 The dataset will be limited to selected words and will be trained according to those words, although it will have an option to expand the dataset according to the needs.

4.7.4 The system will need high processing power, so we will have to take care of those specifications.

4.8 Assumption and Dependencies

4.8.1 Constant power supply

4.8.2 Users must know the limitations in the dataset and that the system will have to be trained for a separate dataset in case they want to make it adaptable to their organization.

4.8.3 The accuracy of the system will have to be considered.

4.9 External Interface Requirements

4.9.1 User interfaces

- **Front-end software**: UI based on Raspbian OS / Windows OS for training.
- **Back-end software**: Python

4.9.2 Display Screen

The main user interface screen will allow the users to monitor the whole process of automation.

4.9.3 Hardware Interface

- Windows Operating System (For training the machine).
- Raspbian OS for Raspberry Pie (For Implementation).

4.9.4 Software Interface

4.9.4.1 Operating System

For training purposes of our system, we have chosen Windows 11 OS for its best support and user-friendliness. And for the end user interface, we will be using Raspbian OS.

4.9.4.2 Database

The user information will be saved in a database / system folder. This data will be in the form of recording of the audio response by the user and will not a permanent record.

4.9.4.3 Python

We have chosen Python as our main programming language as training the machine and all techniques is relatively simpler and more efficient.

4.9.5 Communication Interfaces

Our system runs with Python in the background processing every audio file sent or being listened to real time and a UI based on the Raspbian OS giving real-time feedback.

Chapter 5: ANALYSIS

5 AHIT was aimed at automating the Hearing Impairment test based on Speech Recognition Threshold, in such systems, accuracy has to be taken into account as it is of the primary importance. For this, thorough quality and accuracy check was performed regularly on our dataset so that we get the best possible accuracy.

Here we implemented our Pashto based recognition that contained 15 syllable words. Training and testing parts of dataset were spread into two separate portions. Firstly, features were extracted using MFCC algorithm and then machine-learning algorithms were applied on the data.

For our system, we tried two different algorithms (Naïve Bayesian and CNN) to check which algorithm supports our system the best. However, we couldn't get success in implementation of Naïve Bayesian. By exploring different approaches for our system, we came to know that CNN can be very effective in automatically learning features directly from the dataset during the training process. Here we used MFCC for tuning of CNN, which worked well. The results of our system demonstrate that well trained Convolutional Neural Network maximizes the performance in our system. Our proposed system was shown to improve and add value to previous approaches (as discussed in chapter 2.6).

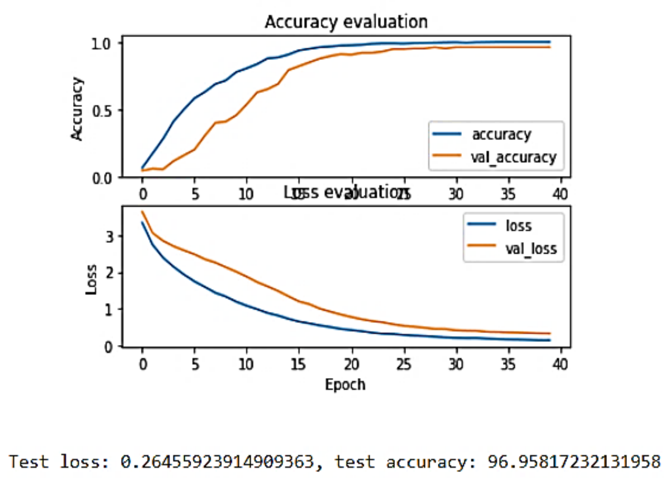

Figure 5.1 Model Results

Chapter 6: Future Work

6 AHIT can be extended further to automate the process of hearing impairment test on large vocabulary. We will try to add more dataset to our project to cover as much Pashto and other local languages as possible. More the data, more accuracy could be achieved from the system. Moreover, to depend on it as an accurate device, we need to cover all the possible vocabulary of Pashto covering all the dialects. For the greater cause, this database will be open to being shared with researchers who want to work on any such system in the Pashto language. As of now, we have completed the model training and predictor coding, but we will try to move it to a standalone device, so that it could be deployed easily in the hospitals and can be used domestically at home.

6.1 SRT Test Result Comparison

The results being obtained by AHIT will be further compared and verified with the results of the same test being performed by the audiologist manually. To verify the practical accuracy of the device.

6.2 Increasing Accuracy

Currently our system is based on 3900 utterances of 15 words by different speakers. The accuracy of our system based on this system using CNN. However, further, to improve this accuracy, different other machine learning algorithms are also needed to be tested.

6.3 Global Domain

We have made a prototype of our idea-based Pushto language and automation of SRT test. However, by introducing a different data set of any other language the device can be made useful for humans with different dialects and cultures. Additionally, introduction of automated Pure Tone Audiometry Test will make this device a global hearing impairment testing device.

6.4 A Patent Standalone Device

As of now, coding of the model training and predictor is done with trial on raspberry pi 4 in process. Since we want to use this application mainly in hospitals, so a standalone device will be preferred to deploy it directly. We will work on making a proper standalone device powered by Raspberry pi 4 and connected to an LCD screen and will apply to get it Patent.

Chapter 7: Conclusion

7 The project "Automated Hearing Impairment Testing using Speech Recognition Threshold based on Automatic Speech Recognition" provides a prototype of a proposed solution for automating the process of hearing impairment testing. This prototype will prove to be an efficient and automated way of the test, resulting in reducing the requirement of manpower as well as the cost effect. The uniqueness of this project i.e., working with the Pashto language has never been catered for in the past. Research to overcome the language barrier while conducting (the test ---) will allow the treatment to reach rural areas of KPK region, which was otherwise not possible. Once implemented on full scale, it will prove to be an important device for upkeep and wellbeing of rural population of KPK region. To our knowledge, no dataset for Pashto words has been collected or worked on, except for counting. Our dataset will be open sourced for future works on the Pashto language. Keeping in view the need for technology in every field, AHIT is a very smart solution for resolving the communication barrier between audiologists and patients by completely automating the whole process. AHIT system can be used in many KPK region of Pakistan and few regions of Afghanistan and by changing the dataset it can be implementable for any spoken language in the world.

References and Work Cited

1. Organization WH, et al. Millions of People in the World have Hearing Loss that can be Treated or Prevented. Awareness is the Key to Prevention. 2013.
2. Organization WH, et al. Deafness and hearing loss. 2015. http:// www.who.int/mediacentre/factsheets/fs300/en/ (visited on 01/16/ 2016). 2014.
3. Martin FN, Champlin CA, Chambers JA. Seventh survey of audiometric practices in the United States. J-Amer Acad Audiol. 1998; 9:95–104.
4. MD R. Audiological survey. 2000.
5. H. Deprez, E. Yılmaz, S. Lievens, and H. Van hamme, “Automating speech reception threshold measurements using automatic speech recognition,” In Proc. 4 th SLPAT Workshop, pp. 35-40, Grenoble, France, Aug. 2013.
6. Abdul Rehman, Luan Jingdong, Imran Hussain, 2015. *Pacific Science Review B: Humanities and Social Sciences*: The province wise literacy rate in Pakistan and its impact on the economy.
7. J. Ashraf, N. Iqbal, N. Khattak and A. Zaidi, "Speaker Independent Urdu Speech Recognition Using HMM", 2021.
8. T. AlHanai, W. Hsu and J. Glass, "DEVELOPMENT OF THE MIT ASR SYSTEM FOR THE 2016 ARABIC MULTI-GENRE BROADCAST CHALLENGE", Groups.csail.mit.edu, 2021.
9. Y. Kumar and N. Singh, "An automatic speech recognition system for spontaneous Punjabi speech corpus", International Journal of Speech Technology, vol. 20, no. 2, pp. 297-303, 2017.

10. A. Arbab Waseem, N. Ahmad and H. Ali, "Pashto Spoken Digits Database for the Automatic Speech Recognition Research", 2012.

11. Ahmed, Irfan & Ali, Hazrat & Ahmad, Nasir & Ahmad, Gulzar. (2012). The development of isolated words corpus of Pashto for the automatic speech recognition research. 139-143. 10.1109/ICRAI.2012.6413380.

12. B. Zada and R. Ullah, "Pashto isolated digits recognition using deep convolutional neural network", Heliyon, vol. 6, no. 2, p. e03372, 2020.

13. S. Khan, H. Ali, Z. Ullah, N. Minallah, S. Maqsood, A. Hafeez KNN and ANN-based recognition of handwritten Pashto letters using zoning features Mach. Learn., 9 (10) (2018)

14. H.B. Chauhan, B.A. Tanawala Comparative study of MFCC and LPC algorithms for Gujrati isolated word recognition Int. J. Innovat. Res. Comput. Commun. Eng., 3 (2) (2015), pp. 822-826 View Record in Scopus.

15. Emre Yılmaz, Joris Peleman, Stefan Lievens, Hugo Van hamme, Cochlear Technology Center, Mechelen, Belgium, ESAT: Speech Reception Threshold Measurement Using Automatic Speech Recognition.

16. Speech Audiometry, March 31, 2020, In "OPHTHALMOLOGY".

17. Shibli Nisar, Muhammad Tariq, Ahsan Adeel, Mandar Gogate, Amir Hussain, Cognitive Computation (2019) 11:489–502: Cognitively Inspired Feature Extraction and Speech Recognition for Automated Hearing Loss Testing, Springer Science + Business Media, LLC, part of Springer Nature 2019.

PLAGIARISM (TURNITN) REPORT

Danish Thesis

ORIGINALITY REPORT

10%	8%	3%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	entokey.com Internet Source	5%
2	Tuka AlHanai, Wei-Ning Hsu, James Glass. "Development of the MIT ASR system for the 2016 Arabic Multi-genre Broadcast Challenge", 2016 IEEE Spoken Language Technology Workshop (SLT), 2016 Publication	2%
3	Submitted to University of Wolverhampton Student Paper	1%
4	Submitted to Higher Education Commission Pakistan Student Paper	1%
5	ieeexplore.ieee.org Internet Source	<1%
6	Bakht Zada, Rahim Ullah. "Pashto isolated digits recognition using deep convolutional neural network", Heliyon, 2020 Publication	<1%
7	Submitted to University of Greenwich Student Paper	

<1%

8	hdl.handle.net Internet Source	<1%
9	www.iapad.org Internet Source	<1%
10	www.semanticscholar.org Internet Source	<1%
11	www.ijcaonline.org Internet Source	<1%
12	Javed Ashraf, Naveed Iqbal, Naveed Sarfraz Khattak, Ather Mohsin Zaidi. "Chapter 14 Speaker Independent Urdu Speech Recognition Using HMM", Springer Science and Business Media LLC, 2010 Publication	<1%
13	Mona A. Azim, Wedad Hussein, Nagwa L. Badr. "Chapter 17 Spoken Arabic Digits Recognition System Using Convolutional Neural Network", Springer Science and Business Media LLC, 2021 Publication	<1%
14	etd.aau.edu.et Internet Source	<1%
15	link.springer.com Internet Source	<1%

16	www.icassp2014.org Internet Source	<1 %
17	Edward Harold Liao, Kim Ganareal, Chirstian Clarence Paguia, Cesar Agreda, Manolito Octaviano, Ramon Rodriguez. "Towards the Development of Automatic Speech Recognition for Bikol and Kapampangan", 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), 2019 Publication	<1 %
18	Yogesh Kumar, Navdeep Singh. "An automatic speech recognition system for spontaneous Punjabi speech corpus", International Journal of Speech Technology, 2017 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		